eric Enteropathogen Resource Integration Center

Bioinformatics Resource Center

Methods for Annnotating Features Other Than Protein-Coding Genes

Guy Plunkett III

BRC4 UAB December 7, 2006

RNA genes & Insertion Sequences

rRNA (ribosomal RNAs)

- BLASTN, individual alignments

tRNA (transfer RNAs)

- tRNAScan-SE plus manual clean-up
  - tRNA-His G at position -1
  - tRNA-Sec longer amino-acyl stem
  - CAU ac (tRNA-fMet, tRNA-Met, tRNA-Ile)

misc_RNA (ncRNA; miscellaneous non-coding RNAs)

- BLASTN, context, Rfam
- Infernal

Insertion sequences

- RepeatMasker and IS Finder

Pseudogenes

A gene that is disrupted in the particular strain or isolate whose genome was sequenced.

Recognized as such by comparison to a related organism where the wild-type or "ancestral" state is seen.

Distinguished from missense mutations, where the gene is still intact but may have altered functionality; in-frame (mod3) indels

Disruption can be due to in-frame stop codons, frameshifts, the insertion of IS elements, prophages, or islands, deletions, and more complex rearrangements.

Pseudogenes

Potential pseudogenes cover a spectrum of cases:

(1) An intact ortholog (allele) occurs in another strain of the same or a closely related species. These are relatively straightforward, and are often resolvable at the nucleotide level.

(2) An intact homolog (ortholog or paralog) occurs in another, more distantly related organism (e.g., another enterobacterial genus). These can usually be resolved at the level of potential protein products.

(3) Partial homology to a gene in another organism. These are unclear, and identification as a pseudogene is open to question.

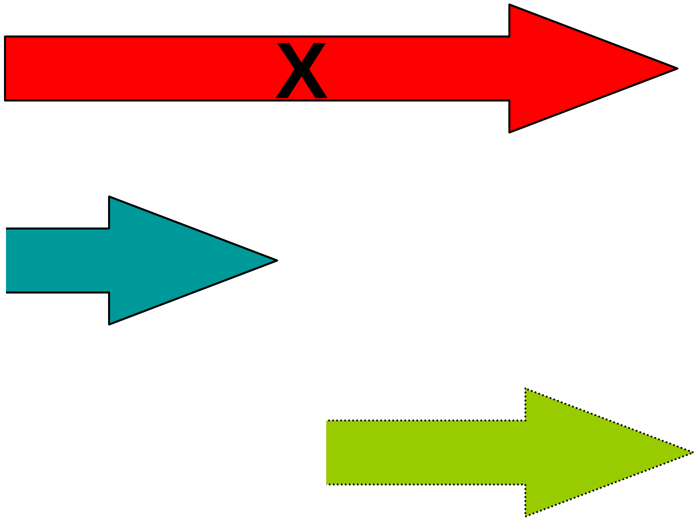
Fused genes; e.g., the two activities of the bifunctional *trpD* in *Escherichia coli* are encoded by distinct genes (*trpD* and *trpG*) in *Yersinia pestis*

Pseudogenes

Is the consequence of a pseudogene something other than the straightforward loss of function?

What to annotate

- **the extent of the pseudogene**
- **the underlying CDS remnants**
- **both**



Pseudogenes

How to annotate

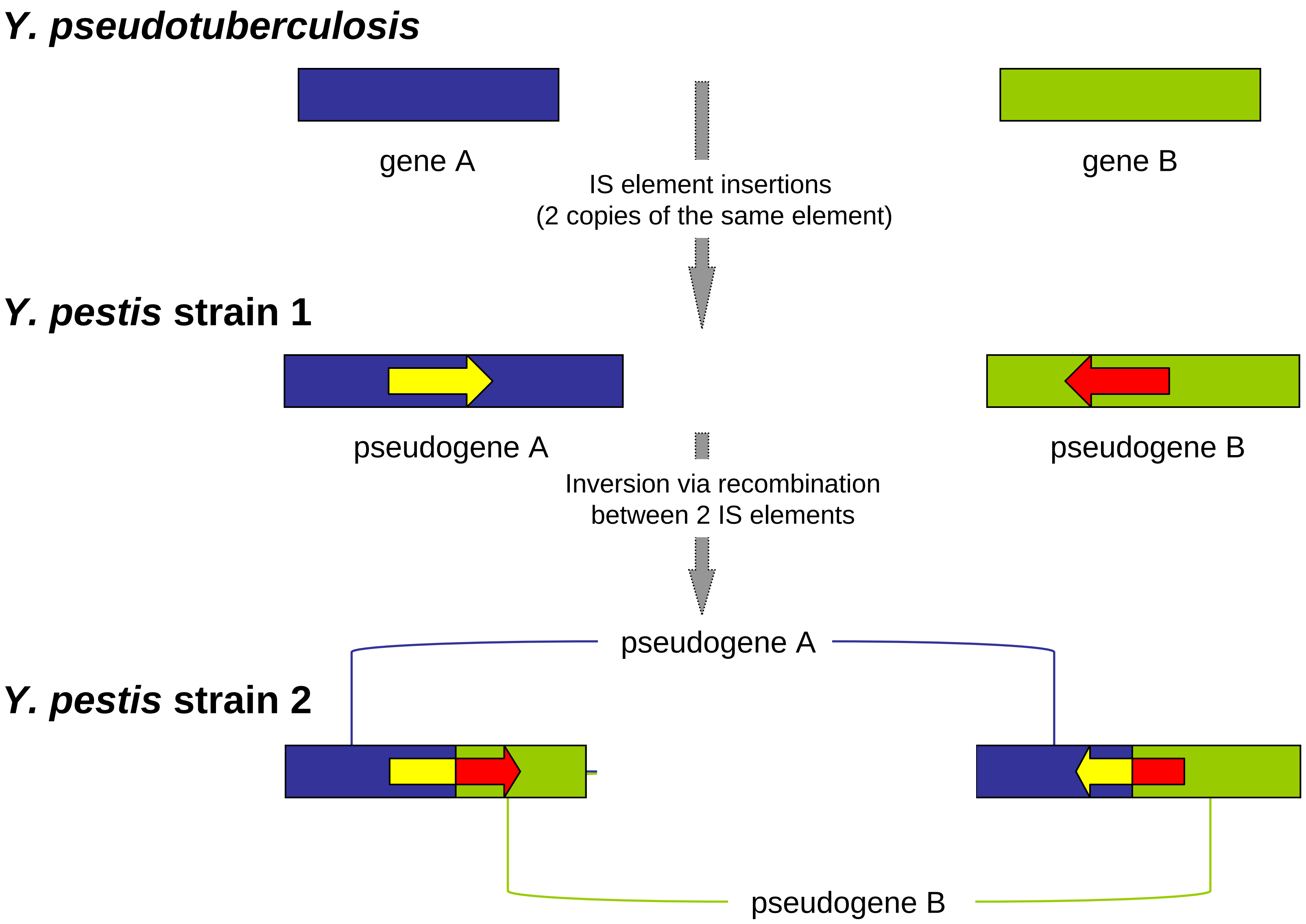
CDS with a /pseudo qualifier
gene with no underlying CDS
misc_feature
-- or not annotated at all!

existing SO term eukaryocentric

To facilitate dealing with them, we have introduced a new feature type of pseudogene within ERIC/ASAP, which can be remapped to any NCBI feature type for a GenBank (re)submission or GFF3 file

Complex situations: pseudogene parts separated by large-scale rearrangements. We have examples where the parts are on different strands, half a genome away!

Pseudogenes -- a complex example



Y. pseudotuberculosis

gene A

gene B

IS element insertions
(2 copies of the same element)

Y. pestis strain 1

pseudogene A

pseudogene B

Inversion via recombination
between 2 IS elements

Y. pestis strain 2

pseudogene A

pseudogene B